DOCUMENT RESUME

ED 159 210 TM 007 516

AUTHOR Lewy, Arieh; Doron, Rina
TITLE Group Tailored Tests and Some Problems of their Utilization.
PUB DATE Jun 77
NOTE 16p.; Paper presented at the International Symposium on Educational Testing (3rd, Leyden, The Netherlands, June, 1977)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Achievement Tests; Criterion Referenced Tests; *Group Norms; *Group Tests; Item Analysis; *Item Banks; Item Sampling; *Norm Referenced Tests; Norms; Statistical Analysis; *Test Construction; Testing; Testing Problems; Test Items
IDENTIFIERS *Tailored Testing

ABSTRACT

The concept of tailored testing for individuals is applied to the construction of tests for special groups and extended to apply to item content as well as item difficulty. It is suggested that evaluators may decide to construct tests on the basis of a unique combination of items drawn from an item bank to fit the need of a particular group. At the same time they may be interested in relating achievements obtained on such tests to some population norms. In order to do this it is necessary to estimate the population mean and standard deviation of each test composed of a unique set of items which, in the past, had not been administered as a single test. This requires finding a way to estimate population standard deviations. The present paper explores the possibility of using two techniques for estimating this parameter; one is Horst's formula for estimating standard deviation and the second one is based on the estimated interitem correlation matrix. In order to examine the adequacy of these estimation formulae, 30 sets of items were selected from the item bank, and for each of these sets the estimated values were compared with the observed ones. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159210

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Arieh Lewy
R. Doron

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

GROUP TAILORED TESTS AND SOME PROBLEMS OF THEIR UTILIZATION

Arieh Lewy and Rina Doron
Tel Aviv University

The Emergence of Tailored Tests

In tailored tests, the examinee answers only a selected set of items out of the total test. Usually one is asked to start with a single or with a block of items of medium difficulty level, and if the score on these items is satisfactory, one is instructed to answer more difficult items, and conversely. if the score is not satisfactory, one is instructed to answer easier items.

Originally tailored tests aimed at increasing the effectiveness of measurement. Tailored tests were introduced to substitute commercially published or standardized tests. Researchers studying the problems of tailored tests concerned themselves mainly with measurement and tried to answer questions such as: how to score response, how to assign items to examinees, etc. A powerful motive for individualizing testing was the aim of obtaining maximum measurement precision with minimum amount of testing time.

Lord (1970) found that the procedure of tailoring tests contributed very little to the precision of the measurement and he expressed doubts as to their merit, at least from a measurement point of view.

TM 005 516

Presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, June, 1977. The Study was supported by a grant from the Ford Foundation received through the Israel Foundation Trustees.

However, in establishing a system for measuring educational outcomes, one has to consider a variety of problems beyond the issue of precision. A global view of achievement testing in school points to certain drawbacks in utilizing nationally produced standardized tests and calls attention to the need to find alternatives to such tests.

One substitue for standardized tests is the utilization of selected sets of items drawn from an item bank and matched to the particular needs of a group of students. Tests of this type can be called group tailored tests. According to this conception, the tailoring serves the purpose of adjusting the test to the needs of a particular group of students and not to the ability level of each individual.

The Drawbacks of Using Standardized Tests

Three problems related to the utilization of standardized tests will be mentioned here: (1) Differences in curricula, (2) Differences in ability level of groups, and (3) Wearing out of standardized test.

1. Differences in Curricula

Most decisions related to selecting objectives and learning contents in today's schools are made on the basis of considering the characteristics of groups rather than on the basis of considering the needs of individuals.

School curriculum is neither individualized nor is it uniform for whole educational systems. Much variation can be observed between teachers from the point of view of emphasizing specific objectives and contents (Gallagher, 1970). Moreover in most national educational systems regional, socio-economic, religious, ethnic, linguistic, etc., differences are associated with differences in school curricula. Consequently educational achievement need not be tested by fully individualized tests, and cannot be fairly tested by test standardized for the whole educational system.

Walker and Shaffarzik (1974) have shown that most achievement tests are biased towards a particular curriculum. Since within a system a variety of curricula may be used, in a comparative survey of educational achievements, some schools (which use curricula different from the one which served as a basis for the test construction) are badly discriminated against.

2. Difference in Ability Level of Groups

Evidence on differences between achievements of students in various schools and various subsystems of schools is presented in the IEA study (Thorndike, 1973). According to the IEA[1] results, the median value of the percent of variance due to between-school differences is about 30 percent of the total variance in developed countries and about 60 percent in developing countries.

---

[1] IEA, International Association for the Evaluation of Educational Achievement, University of Stockholm, Department of International Education.

Owing to substantial differences in the level of general achievement between schools and between educational subsystems, nationally standardized tests can prove to be too difficult or too easy for a certain group of students taking a test. Giving a very difficult test to a low-achieving group may have serious social and educational consequences. Thus, in constructing tests, precision of measurement should not be the only consideration. One must be aware of the differences betwen a situation in which a test does not fully fit the needs of certain individuals' learning in a group and situation in which a test is unsuitable for measuring achievements within a whole subgroup of a population. While in the first case, one may or may not recommend the use of individually tailored tests, it seems that in the second case differentiation between tests given to various groups is an absolute necessity, and, hence, the justification for introducing the notion of group tailored tests.

3. Wearing Out of Tests

Standardized tests can be efficiently used only if the students and teachers are not familiar with the items included in the test. The confidentiality of tests and secrecy of their contents constitute a necessary condition for the validity of the scores. The problem of test security constitutes a vexing problem for those responsible for the management of testing projects (Lippey, 1973). While in the United States, due to the existence of a wide variety of tests in most fields of study, the chances are small that students will familiarize themselves with the contents of the tests; in a small educational system, the situation is different.

Ie the United States, an age-group consists of approximately four million students, In Israel an age-group consists of 50,000 students only, and in many other countries (like Cyprus, Iceland, Liberia, Togo, Zambia) an age-group hardly reaches 15,000. In small countries there are no standardized tests in most fields of studies, let alone alternative ones. The development of a new test is often an event which arouses much interest in educational circles. Under such circumstances tests quickly become overused and items wear out.

The utilization of a unique combination of items for each test-taking occasion is another aspect of test-tailoring which may improve the measurement of educational outcomes.

Item Bank as a Source for Group Tailored Tests

The construction of tailored tests from an item bank entails four steps:

1. Compiling an initial set of items.
2. Continuous up-dating of the item bank.
3. Service.
4. Follow-up.

1. Compiling an initial set of items.

The first step in developing an item bank is compiling an inital set of items related to a series of educational objectives and learning contents. The item bank should contain information about some characteristic features of the items. Thus, for example, an initial set of items produced in Netherlands (Timmer, 1972) contains information about the following characteristics of the item.

a. Objective and content covered by the item.

b. Reference to page numbers in different textbooks which contain learning materials related to the item.

c. Proportion of students in various subpopulations selecting a given response alternative.

d. Point biserial correlation with a specified set of items.

In an item bank constructed in Israel (Doron, 1976), the "wit-value" of each item derived from a sample-free caliberation of the items according to the Rasch model (Rasch, 1960) has been added to the item characteristics listed before.

2. Continuous Up-Dating

The efficient use of an item bank in any educational system requires continuous updating. New items should be added to the bank to cover additional objectives and contents, and to provide alternative items for the existing one in order to prevent the wearing out of the bank. Parameters of the existing items should be updated. Thus, for example, its relatedness to new learning materials produced in the system, changes in its difficulty level, the extent of its use in various schools, etc., should be recorded in the item bank.

3. Service

It is not enough to keep a well organized item bank on computer tapes, in neatly arranged files or in well organized commercially distributed kits. Effective use of an item bank requires that users obtain professional services

in compiling tests, in making valid inferences and sound decisions on the basis of the test results. At least three issues should be considered while selecting items for testing achievements of a particular group of students.

a. Relatedness to the curriculum.

b. The general ability level of the group of learners.

c. Previous encounters with specific set of items.

This third point means that the schools which year after year utilize an item bank should avoid the repeated use of the same items.

4. Follow-up

Finally, continuous follow-up should be made about the utilization of item banks for group tailored tests. One should examine the kind of difficulties which emerge during the process of utilizing the item bank, and measures should be taken to eliminate these difficulties. This can be done through training sessions for users, better manuals, etc.

Norms for Group Tailored Tests

Undoubtedly, the widespread utilizaiton of tailored tests will give new directions for the psychometric research and theory. One may assume that need will be felt to introduce new parameters for characterizing tests and also new computational procedures to obtain estimates for parameters used in the contemporary test theories.

One problem of this type is the estimation of norms for tailored tests. In other words, is it possible to store information which enables an estimate of national norms for any unique combination of items drawn from an item bank? The Rasch model of sample-free test calibration produces some statistics which enables the comparison of groups on the basis of different tests, provided they contain a common set of anchor items. But these statistics do not permit the making of inferences concerning the relative position of a group or of an individual in a defined population. To estimate norms one needs two statistics: the sample mean and its standard deviation. The first of these two statistics can easily be estimated on the basis of the item difficulty indices scored in the item bank. The estimation of the standard deviation is more complicated, since it requires some estimate of the covariance of items appearing in a particular test. To provide an estimate of the standard deviation of a unique set of items, two computational procedures were examined and the following section of this paper contains some empirical data about the adequacy of these estimates. The two methods of estimates are:
(1) Estimating association between items on the basis of item difficulty, and (2) Estimating the variance on the basis of item difficulty.

1. Estimating Association Between Items on the Basis of Item Difficulty

It is practically impossible to produce empirical estimates of the degree of association of all items appearing in an item bank. Even if the bank contains 19,900 values. But in practice, item banks contain thousands of items accumulated during a long process of administrating relatively short tests to different samples of students and therefore the computation of the correlation coefficients becomes impossible.

Nevertheless, in all cases some correlation coefficients among subsets of items are available. These statistics should be recorded in the item bank. If one want to estimate the variance of a set of items selected from the item bank on the basis of the correlation among them, one has first to find out which correlation coefficients are listed in the bank and which are not. Then, the missing values can be estimated in some convenient and efficient way.

In the present study the estimation was based on the assumption that the correlation coefficients between homogeneous sets of items are functions mainly of the item difficulty levels.

To produce such estimates, 10,000 correlation coefficients were computed and ordered according to the difficulty levels of items being correlated. Items were grouped in 20 groups according to intervals of difficulty where a scale of 5 percent constituted one interval. Thus, for example, the difficulty level of 11-15 percent or 16-20 percent constituted a single interval. In this way, a symmetric matrix of 20 x 20 was constructed and the correlation coefficients were assigned to the lower triangular cells of the matrix. Then, the medium value of the correlation coefficients recorded in each particular cell was used as an estimate of correlation for pairs of items with corresponding items difficulty levels (Lewy and Doron, 1976). It turned out that the matrix of estimated correlations had a pattern of simplex (Guttman, 1954) in so far as the high values were close to the principal diagonal trace of the matrix. The actual estimates did not exceed the value of 0.2.

The estimated standard deviations are listed in Table 1.

Table 1: Observed and Expected Standard Deviations

The Test	No. of Items	Observed SD	Estimated by Correlations	Difference	Estimated by Item Difficulty	Difference
A	11	3.084	2.879	.205	3.624	-.540
B	13	2.245	2.232	.013	2.779	-.534
C	8	1.720	1.769	.049	2.298	-.578
D	9	2.503	2.188	.315	2.491	.012
E	9	1.910	1.828	.082	1.572	.337
F	9	2.099	1.659	.412	2.337	-.238
G	10	2.027	1.659	.368	2.616	-.589
H	13	2.902	2.715	.187	3.545	-.643
I	8	2.271	2.025	.246	2.455	-.184
J	10	3.309	2.135	1.174	2.905	.404
K	12	3.100	2.511	.589	3.537	-.437
L	12	3.043	2.596	.447	3.649	-.606
M	8	2.122	1.760	.362	2.397	-.275
N	12	2.875	2.407	.468	3.181	-.306
O	10	2.250	2.210	.040	3.113	-.863
P	13	3.879	2.415	1.461	3.917	-.038
Q	10	3.560	2.495	1.066	3.481	.079
R	9	2.743	2.055	.688	3.012	-.269
S	12	3.153	2.938	.215	3.822	-.669
T	6	1.900	1.553	.348	1.714	.187
U	8	2.575	1.901	.678	2.510	.065
V	8	2.419	2.128	.291	2.478	-.058
W	7	2.958	1.935	1.023	2.389	.569

It can be seen that the estimates, except one, are lower than the observed values. The median discrepancy between the estimated and observed values is approximately 20% of the observed standard deviation. The relatively high discrepancy values between the estimates and the observed values appear mainly in tests of low difficulty level. Examination of the data revealed that the observed values of the correlation between items having 80 percent or 90 percent correct responses are generally higher that their estimated values.

2. Estimating the Variance on the Basis of Item Difficulty Indices

A second attempt to estimate the test variance was based on a formula developed by Horst (1966) and which contained only item difficulty indices. Horst's formula has the following form:

$$s^2 = 2(1-c)\sum_{i=1}^{m} ip_i + \bar{x}(2mc-1-\bar{x}) - cm\,(m-1)$$

Where the items are arranged in descending order of difficulty level:

S^2 = estimated variance

c = probability of guessing

p_i = proportion of person answering correctly item i

m = number of items

$\bar{x}$ = mean of the test.

The rationale of this estimation formula is explained in Horst's work and will not be reported here. In order to obtain empirical evidence about the accuracy of this formula, the estimates of 23 subsets of items were compared with their actual computed values. Table 1 contains the two sets of values.

It can be seen that in many cases, the differences had positive values and in many cases, had negative values. The maximum difference between the observed and estimated value was 38 percent of the actual standard deviation, the median deviation in terms of the percent of the actual standard deviation was 13. It should also be noted that no correlation exists between the errors of the estimates obtained by the two methods described above. In the face of these results, one may recommend the use of these estimates for assigning a group to a particular quartile of a population, but on the basis of this formula, no more accurate estimates can be made with an acceptable degree of precision.

Summary

Within most national educational systems, wide differences exist between various subsystems of schools and between individual teachers from the point of view of the curriculum taught in classes. But within a single class, teaching is still directed to the group and is not fully individualized. Consequently, achievement testing should consider group differences. It is suggested that achievement tests should be group tailored; i.e., they should consider differences in materials taught in schools, but they need not be fully individualized. Group tailored tests should substitute nationally used standardized tests.

Item banks can be used as a source for group tailored tests. The development of item banks used for such purposes entails four steps:

1. Compiling an initial set of items
2. Up-dating
3. Service for users.
4. Follow-up.

While most studies dealing with tailored tests focus on the precision of individual scores, the present study addresses itself to group statistics of tailored tests. It explores the possibility of estimating national norms for group tailored tests on the basis of test item characteristics stored in the item bank. Two methods of estimation were examined.

1. Estimating the test variance on the basis of estimating inter-item correlations.
2. Estimating test variance on the basis of item difficulty indices.

It has been demonstrated that the estimations are crude approximations of the actual parameters, neverthless, they can be used with a sufficient level of accuracy for placing groups or individuals into a certain quartile of the population. It should be noted that in a great many educational studies even such crude information about the achievements of students studying in a particular school or school district has practical significance.

REFERENCES

Doron, R. An item bank in arithmetic, common fractions. Tel Aviv University, 1977 (mimeo. in Hebrew).

Gallagher, J. I. Three studies of the classroom in R. Stake (Ed.), Classroom Observation AREA Monograph Series, 1970, 6, Chicago: Rand MacNally.

Guttman, L. A new approach to factor analysis? The Radex. In Lazarsfeld, P. F. (Ed.), Mathematical thinking in social sciences. Glencoe, Ill.: The Free Press, 1954.

Horst, P. Psychological measurement and prediction. Belmont, Calif.: Wadsworth Publishing Co., 1966.

Lewy, Arieh, & Doron, R. Utilization of test item pools for constructing educational measurement instruments. Unpublished manuscript, Tel Aviv University, 1976.

Lippey, Gerald. The computer can support test construction. Educational Technology, 1973, 10-12.

Lord, F. M. Some test theory for tailored testing. In W. H. Holtzman, (Ed.), Computer-assisted instruction, testing and guidance. New York: Harper and Row, 1970.

Rasch, G. Probalistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Thorndike, R. L. Reading comprehension education in fifteen countries. Stockholm: Almquist and Wiksell, 1973.

Trimmer, J. Vier Keuzeüragen Wiskunde (Dutch). Arnheim: CITO, 1972

Walker, D. F., & Shefferzik, J. Comparing curricula. Review of Educational Research, 1974, 44, 83-111.